Skeleton Filter

Skeleton Filter: A Self-Symmetric Filter for Skeletonization in Noisy Text Images

Xiuxiu Bai, *Member, IEEE,* Lele Ye, Jihua Zhu, Li Zhu, and Taku Komura

***Abstract*—Robustly computing the skeletons of objects in natural images is difficult due to the large variations in shape boundaries and the large amount of noise in the images. Inspired by recent findings in neuroscience, we propose the Skeleton Filter, which is a novel model for skeleton extraction from natural images. The Skeleton Filter consists of a pair of oppositely oriented Gabor-like filters; by applying the Skeleton Filter in various orientations to an image at multiple resolutions and fusing the results, our system can robustly extract the skeleton even under highly noisy conditions. We evaluate the performance of our approach using challenging noisy text datasets and demonstrate that our pipeline realizes state-of-the-art performance for extracting the text skeleton. Moreover, the presence of Gabor filters in the human visual system and the simple architecture of the Skeleton Filter can help explain the strong capabilities of humans in perceiving skeletons of objects, even under dramatically noisy conditions.**

***Index Terms*—skeleton detection, filter, noisy text images.**

I. Introduction

In the visual cortex of the brain, according to theoretical and neuroscience results, skeleton (or medial axis) representation exists in the inferotemporal cortex (IT) [1], [2], [3], [4], [5], [6]. The skeleton is a feature that can well describe the shapes of objects in low dimensions, even highly complex shapes [7]. This representation has substantial benefits for invariant shape coding and parts-based structures and has been widely used in computer vision [8], [9], [10], [11], [12]. For example, skeleton-based matching is more robust to geometric variations compared with shape-based matching [13]. Moreover, skeleton representation performs especially effectively in text recognition [14].

From the perspective of computational modelling, many skeleton extraction approaches have been proposed, which include distance transforms [15], thinning [16], Voronoi diagrams [17], bone graphs [18], and the appearance medial axis transform (AMAT) [19]. A central challenge in the computation of skeletons is instability: the skeleton computation is highly sensitive to boundary noise and variations, namely, small perturbations to the shape cause the emergence/disappearance of branches in the skeleton [20].

From the neuroscience perspective, determining how neural signals encode the information from lower visual signals into a skeleton remains an open problem [1]. According to several neuroscience studies, orientation selectivity exists in the V1 cortex [21], [22]. There is theoretical and experimental evidence that orientation selectivity in the V1 cortex is encoded using a similar pattern to Gabor filters [23]. A Gabor-like filter [24], [25], with a simplified Gabor filter structure, consists of a positive Gaussian filter and a negative Gaussian filter. There is an evidence that human brains conduct operations that are similar to Gaussian filters [26]. How humans encode the orientation information into skeleton features needs to be verified in neuroscience.

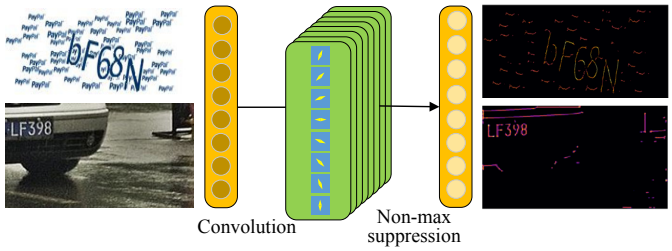


Fig. 1. Skeleton Filter can directly perceive skeletons via convolution and non-max suppression operations.

Inspired by the recent findings in neuroscience that are discussed above, we propose the Skeleton Filter, which is a novel model for skeleton extraction from noisy text images. The Skeleton Filter combines a pair of oppositely oriented Gabor-like filters; by applying Skeleton Filters with various orientations to an image, our model can robustly detect the skeleton even under highly noisy conditions. We use challenging noisy text datasets to evaluate the performance of our method, on which state-of-the-art performance is realized.

Our key contribution is the use of known neuroscience evidence to construct a zero-sum self-symmetric operator model, namely, the Skeleton Filter, which can perceive the skeleton (as in Figure 1) through only a combination of convolution and non-max suppression. Moreover, our model's simplicity and robustness in skeleton detection may describe how the human visual system perceive skeleton features.

The remainder of this paper is organized as follows: Section II introduces related works. Section III introduces the proposed method in detail. In Section IV, we present the experimental results. Finally, we present the conclusions of this work in Section V.

II. Related Work

Since the concept of skeleton representation of 2D shapes was introduced by Blum [3], researchers have developed a se-

Manuscript received February 3, 2019; revised August 13, 2019; accepted September 26, 2019. This work was supported by the National Natural Science Foundation of China under Grants 61802297 and 61972312. (Corresponding author: Xiuxiu Bai.)

X. Bai, L. Ye, J. Zhu and L. Zhu are with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: xiubai@xjtu.edu.cn; yeler082@stu.xjtu.edu.cn; zhujh@xjtu.edu.cn; zhuli@xjtu.edu.cn).

T. Komura is with School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, United Kingdom (e-mail: tkomura@ed.ac.uk).

ries of approaches for efficient and robust skeleton extraction.

Classical methods There are three classical skeletonization techniques: using a distance transform [15] to calculate ridges in a distance map of the boundary points, calculating the skeleton via thinning [16] by layer erosion, and Voronoi-based skeleton extraction [17], which utilizes the Voronoi diagram of the boundary points. Siddiqi et al. [27] propose shocks as the singularities of curve evolution boundaries, which are combined to form an acyclic directed shock graph. A bone graph [18] identifies the ligature structure and restores the non-ligature structures as a skeleton as well as offers improved stability and an intuitive representation of an object's parts.

Most previous skeleton simplification approaches approximate the medial axis via the union of medial spheres and use a local or global threshold. Feldman et al. [11], [28] introduce the Bayesian probability method, which grows the skeleton via a random generation process into the final object skeleton. Garland et al. [29] propose a surface simplification method that uses the iterative contraction of vertex pairs and a quadratic error metric (QEM) to approximate the surface error to generate an approximation of the polygon model. Spherical QEM [30] extends the QEM to a simplified volume by computing the squared distance from a sphere to the containing planes of its associated boundary triangles. The angle-based filtering method [31] computes the angle that is formed by each medial axis point and its two closest points on the shape boundary. The λ-center axis method [32] uses the circumference of the closest point to the midpoint as the clipping criterion.

In 3D scenes, the scale axis transformation (SAT) [33] utilizes the spatially adaptive classification of geometric features to generate internal representations at various levels of abstraction. Faraj et al. [34] propose a progressive central axis simplification via the continuous edge folding of the input center axis hierarchy. Sun et al. [35] propose a Hausdorff-error-based method by computing a volume approximation for medial axis simplification. Li et al. [20] use quadratic error minimization to compute an accurate linear approximation of the skeleton.

Unsupervised methods Lindeberg [36] defines the skeleton as the points at which the intensity attains its local maximum or minimum in the main eigendirection of the Hessian matrix. Jang et al. [37] extract the skeleton by calculating the pseudo distance map from the edge-strength function using a partial differential equation. Yu et al. [38] extract the skeleton from a skeletal intensity map that is calculated from a diffuse vector field to provide a measure of the likelihood of each pixel on the skeleton. Direkoglu et al. [39] extract skeletons from grayscale images based on anisotropic thermal diffusion analogies and use the skeleton strength map to describe the likelihood of a point being part of the skeleton. Mignotte [40] quantifies the skeleton between each pair of line segments via a Hough-style voting approach and an averaging procedure to remove noise. Tsogkas et al. propose AMAT [19], which regards the detection of the skeleton of a natural image as a weighted geometric set cover problem.

As an unsupervised method, in contrast to the methods that are described above, our Skeleton Filter possesses the zero-sum self-symmetric structure. Therefore, it requires only convolution and non-max suppression operations to directly identify the skeleton points.

Supervised learning methods Tsogkas et al. [41] extract the symmetry structure by using multiple instance learning to combine cues, such as texture, color, structure, and spectral clustering information. Widynski et al. [42] formulate the symmetry detection problem as a spatial Bayesian tracking task using a sequential Monte Carlo method and an adaptive semi-local geometric model. Recently, a series of deep learning approaches for extracting skeletons, such as deep skeleton [43], side-output residual network (SRN) [44] and Rich SRN (RSRN) [45]. RSRN fuses side-outputs in a deep-to-shallow manner to decrease the residual between the detection result and the ground-truth [45]. These deep learning approaches require large training datasets and ground-truth skeletons of images. Since there are no ground-truth skeleton in the text datasets, such approaches are not applicable to this problem. In this paper, we focus our comparisons on unsupervised skeleton extraction methods.

In summary, the previous approaches are sensitive to large variations in the shape boundaries and to noise in the images. Moreover, these approaches are difficult to explain from the perspective of neuroscience.

III. Skeleton Filter

A. Neuroscientific basis

In neuroscience, fundamental evidence supports the existence/establishment of the skeleton representation within the brain.

1) At the neural level, there is evidence of skeleton signals in the V1 [46] and IT visual cortices [1]. IT activities embody a basis set for simultaneously representing the skeletons and the external shapes of complex objects [1];

2) Orientation features in the V1 visual cortex are encoded using a similar pattern of Gabor filter [23];

3) The Gaussian filter is a well-established model of neurons in the visual cortex [26];

4) Convolution and non-max suppression operations are utilized in the responses of visual neurons [47].

The determination of how the brain encodes low-level information into a skeleton feature remains an open problem in neuroscience. Based on these neuroscience evidence, we propose a novel model for direct skeleton detection.

B. Principles

Inspired by the above recent findings, we propose the Skeleton Filter that consists of a pair of oppositely oriented Gabor-like filters [24], each of which is composed of a positive and a negative isotropic Gaussian filter. A Gabor filter is a Gaussian kernel function that is modulated by a sinusoidal plane wave. In practice, a Gabor-like filter [24], [25], a simplified structure of a Gabor filter, which consists of a positive and a negative isotropic Gaussian filter, and is used to detect orientations and edges.

Figure 2 visually illustrates the structures of a 1D Gabor-like filter and the Skeleton Filter. A pair of oppositely oriented 1D

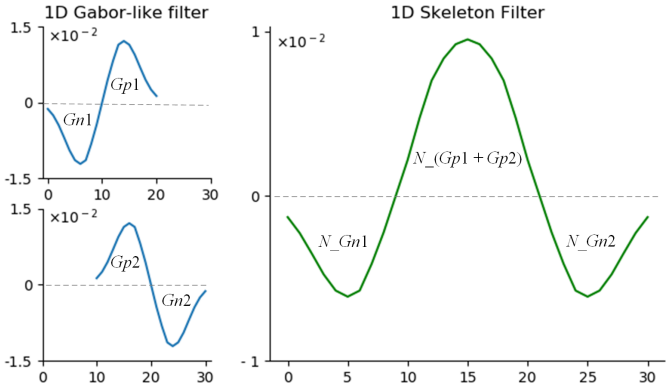


Fig. 2. The structure of the 1D Gabor-like filter and the Skeleton Filter. The 1D Gabor-like filter consists of a positive Gaussian filter (*Gp*) and a negative Gaussian filter (*Gn*). The standard deviations of *Gp* and *Gn* are the same. The 1D Skeleton Filter is constructed by a pair of oppositely oriented 1D Gabor-like filters. The formation process of the 1D Skeleton Filter is as follows: 1) the positive Gaussian filters in the center are summed (*Gp*1 + *Gp*2); 2) the absolute value of the whole 1D Skeleton Filter is unified normalized (*N*(*Gn*1, *Gp*1 + *Gp*2, *Gn*2)). In the 1D Skeleton Filter, the sum of the positive and negative values is zero.

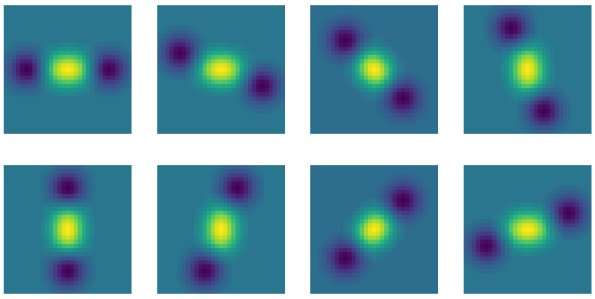


Fig. 3. Skeleton Filter banks for detecting 8 oriented skeleton in an input image. The light green filter is a positive Gaussian filter and the dark blue filter is a negative Gaussian filter. The Gabor-like filter combines a positive and a negative isotropic Gaussian filter. Skeleton Filters consist of a pair of oppositely oriented Gabor-like filters. Skeleton Filters possess three types of self-symmetric properties: reflection symmetry, rotational symmetry, and internal (positive value) and external (negative value) symmetry.

Gabor-like filters are combined to form a 1D Skeleton Filter by applying summation and normalization operations.

Figure 3 shows a bank of Skeleton Filters. The positive Gaussian filters are in the center of each filter and the negative Gaussian filters are in the peripheral. The formation process is the same for 1D and 2D Skeleton Filters. The positive Gaussian filters in the center are summed and the whole Skeleton Filter is unified normalized. In the Skeleton Filter, the sum of the positive and negative values is zero.

Next we use mathematical concepts to formally construct our approach. The 2D Gaussian filter is expressed as

$$g(x, y, \mu_1, \mu_2) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu_1)^2+(y-\mu_2)^2}{2\sigma^2}} \tag{1}$$

where x denotes the distance from the origin along the horizontal axis, y denotes the distance from the origin along the vertical axis, σ denotes the standard deviation, and μ denotes the expected value.

The 2D Gabor-like filter is expressed as

$$\begin{aligned} &gab_0(x, y) \\ &= \begin{cases} -g(x, y, \frac{w}{2}, \frac{w}{2}), 0 \leqslant x < w, 0 \leqslant y \leqslant w \\ g(x, y, \frac{3w}{2}, \frac{w}{2}) - g(x, y, \frac{w}{2}, \frac{w}{2}), x = w, 0 \leqslant y \leqslant w \\ g(x, y, \frac{3w}{2}, \frac{w}{2}), w < x \leqslant 2w, 0 \leqslant y \leqslant w \end{cases} \\ &= \begin{cases} -g(x, y, \frac{w}{2}, \frac{w}{2}), 0 \leqslant x < w, 0 \leqslant y \leqslant w \\ 0, x = w, 0 \leqslant y \leqslant w \\ g(x, y, \frac{3w}{2}, \frac{w}{2}), w < x \leqslant 2w, 0 \leqslant y \leqslant w \end{cases} \end{aligned} \tag{2}$$

$$\begin{aligned} &gab_1(x, y) \\ &= \begin{cases} g(x, y, \frac{3w}{2} + 1, \frac{w}{2}), w + 1 \leqslant x < 2w + 1, 0 \leqslant y \leqslant w \\ g(x, y, \frac{3w}{2} + 1, \frac{w}{2}) - g(x, y, \frac{5w}{2} + 1, \frac{w}{2}), x = 2w + 1, \\ \qquad 0 \leqslant y \leqslant w \\ -g(x, y, \frac{5w}{2} + 1, \frac{w}{2}), 2w + 1 < x \leqslant 3w + 1, 0 \leqslant y \leqslant w \end{cases} \\ &= \begin{cases} g(x, y, \frac{3w}{2} + 1, \frac{w}{2}), w + 1 \leqslant x < 2w + 1, 0 \leqslant y \leqslant w \\ 0, x = 2w + 1, 0 \leqslant y \leqslant w \\ -g(x, y, \frac{5w}{2} + 1, \frac{w}{2}), 2w < x \leqslant 3w + 1, 0 \leqslant y \leqslant w \end{cases} \end{aligned} \tag{3}$$

where w denotes the width. In gab_0, the distribution uses the negative Gaussian filter if $0 \leqslant x < w$; the distribution uses the positive Gaussian filter if $w < x \leqslant 2w$; and the positive and negative Gaussian filters are overlapped if $x = w$. The formation process of gab_1 is similar to that of gab_0. gab_0 and gab_1 can detect the opposite orientation of the shape.

The 2D Skeleton Filter is defined as follows:

$$\begin{aligned} &s(x, y) \\ &= gab_0(x, y) + gab_1(x, y), 0 \leqslant x \leqslant 3w + 1, 0 \leqslant y \leqslant w \\ &= \begin{cases} -g(x, y, \frac{w}{2}, \frac{w}{2}), 0 \leqslant x < w, 0 \leqslant y \leqslant w \\ g(x, y, \frac{3w}{2}, \frac{w}{2}) + g(x, y, \frac{3w}{2} + 1, \frac{w}{2}), w < x \leqslant 2w, \\ \qquad 0 \leqslant y \leqslant w \\ 0, x = w \text{ and } x = 2w + 1, 0 \leqslant y \leqslant w \\ -g(x, y, \frac{5w}{2} + 1, \frac{w}{2}), 2w + 1 < x \leqslant 3w + 1, 0 \leqslant y \leqslant w \end{cases} \end{aligned} \tag{4}$$

The 2D Skeleton Filter is unified normalized.

$$s_{norm}(x, y) = norm(s(x, y)), 0 \leqslant x \leqslant 3w + 1, 0 \leqslant y \leqslant w \tag{5}$$

The pattern of Eq. (5) corresponds to the first example in Figure 3. The other oriented filter banks can be obtained by rotating the first example in the xy-plane.

$$s^{(i)}_{norm} = rotate(s_{norm}, \frac{2\pi}{k} i), i = 0, 1, ..., k - 1 \tag{6}$$

where k is the number of filter banks of Skeleton Filters.

Based on the above definition, a bank of Skeleton Filters contains three types of self-symmetric properties: reflection symmetry, rotational symmetry, and internal (positive value) and external (negative value) symmetry. From the theoretical perspective, this zero-sum self-symmetric property could maintain its equilibrium state, which increases the possibility

of a receptive field with this structure (Skeleton-like Filter) existing in the visual cortex.

Figure 4 illustrates the principle of Skeleton Filter detection. The Skeleton Filter has a self-symmetric structure. Each simplified Gabor filter responds to edges that are vertical relative to its orientation, namely, the direction of the arrow in Figure 4. The two opposite orientation vectors are vertical to the line that connects the origins of the Gabor-like filters; thus, the central point of the line is the medial axis position.

Algorithm 1 outlines the detailed skeleton detection procedure that uses the Skeleton Filter. To compute the skeleton of an object, we can apply convolutional operations to the input image with the Skeleton Filter banks and fuse them to compute the skeleton feature map. We apply non-max suppression operations to the bottom-up feature messages, such that only a single oriented skeleton is active at each pixel. To extract the skeletons of objects with different scales, we apply the Skeleton Filters to an image at multiple resolutions and fuse the results. Figure 1 presents an overview of the process of extracting the skeleton from an image using the Skeleton Filter.

From the theoretical perspective, the reasons for robustly detecting skeletons under noisy conditions are as follows:

1) A Gaussian filter can moderately smooth the image, thereby reducing the effects of obvious noise and removing spiky edges.

2) The noise distribution in an image typically does not have a symmetric structure; however, our Skeleton Filter which has a self-symmetric architecture, is only sensitive to the symmetric parts of the images. Hence, it can filter out this type of noise distribution.

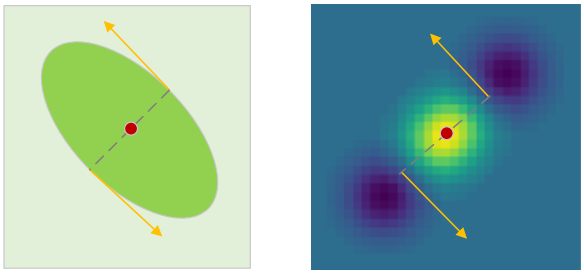
3) In various cases, the noise distribution is relatively uniform, such as in rainy conditions. In our Skeleton Filter, the sum of the positive and negative values is zero. Hence, it will be equal to zero when a uniform noise distribution is convolved with the Skeleton Filter. Therefore, it can filter out this type of uniform noise distribution.

In summary, the zero-sum self-symmetric architecture with Gaussian filters results in the robust extraction of the skeleton, even under highly noisy conditions.

Algorithm 1 Skeleton detection by Skeleton Filter (SF)

Input: Image X, Skeleton Filter SF, number of filter banks k
Output: skeleton Y

Compute $s_{norm}^{(i)}$ of SF by Eq. (6)
// Apply FFT convolution operation to filter X
for each filter bank i in $s_{norm}^{(i)}$ **do**
  $filtered[i]$ = fftconvolve(X, $s_{norm}^{(i)}$)
end for
// apply non-max suppression to the $filtered[i]$, so that only a single orientated skeleton is active at a pixel
for each filter bank i in k **do**
  $suppressed[i]$ = non_max_suppression($filtered[i]$)
end for
// combine $suppressed[i]$ to a unified map M
M = combined ($suppressed[i]$)
Y = threshold (M)
return Y



(a) Object example (b) One orientation of Skeleton Filter

Fig. 4. Skeleton Filter detection principle. Each Gabor-like filter can detect oriented lines in the direction of the arrows. Two opposite orientations are vertical to the line connecting the center of the two Gabor-like filters. This configuration produces a medial axis at the red point when the filter is convolved with an object that has the same orientation as the Skeleton Filter.

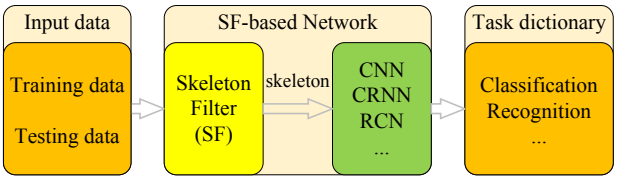


Fig. 5. SF-based network. The Skeleton Filter serves as a preprocessing module which obtains the skeleton results.

Neuroscience research has proven that a scheme based on similar Gabor filters that can perceive orientation features exists in the human visual V1 cortex. A suitable organization of Gabor filters can potentially perceive the skeleton feature in the V1 and IT cortices, although no configuration that can induce the skeleton has been identified in previous works. The Skeleton Filter introduced in this study can naturally propagate local orientation features into a global skeleton feature. Thus, the Skeleton Filter could potentially be the configuration of the Gabor-like filters in the human visual system for recognizing skeleton features.

C. Application

The Skeleton Filter can play an important role in various tasks. We propose a skeleton-filter-based network (SF-based network), which uses Skeleton Filters as a preprocessing module, illustrated in Figure 5. The networks can include, for example, convolutional neural network (CNN), a recursive cortical network (RCN) [24], or a convolutional recurrent neural network (CRNN) [48]. In this paper, we design a Skeleton Filter CNN (SF-CNN) for conducting the classification task. SF-CNN uses the Skeleton Filter to obtain the skeletons of input images and applies CNN for classification. Similarly, we design a Skeleton Filter CRNN (SF-CRNN) to conduct the recognition task. SF-CRNN uses the Skeleton Filter to obtain the skeletons of input images and applies CRNN for recognition.

D. Time complexity

Our Skeleton Filter uses convolution and non-max suppression operations. The most time-consuming step is the convolution operation, which applies the FFT operation. The time complexity of FFT is $O(nlog(n))$, where n is the data size. In our method, the time complexity is $kO(nlog(n))$, where k denotes the number of filter banks. The procedure can be easily parallelized by simultaneously calculating k channels. In our system, k is set to eight. Our non-parallelized Python implementation on an Intel Xeon E3-1505M CPU takes about 0.07 seconds on a 200 $\times$ 200 gray image.

IV. Experiments

A. Experimental Setup

Datasets We evaluate the performance of our approach on several challenging noisy text datasets: CAPTCHA, Noisy MNIST, SVHN, ICDAR 2013, and IIIT 5k-word (IIIT5k). **CAPTCHA** (Completely Automated Public Turing test to tell Computers and Humans Apart) is a type of challenge test for determining whether the user is a human or a computer [49]. CAPTCHAs are typically used by websites to block automated interactions. We evaluate our approach on these CAPTCHA datasets inlcuding **BotDetect**, **Paypal** and **reCAPTCHA**. These datasets include text images with various noises and distortions that make automatic recognition difficult. **Noisy MNIST** [24], which differs from MNIST, includes six variants of noise such as background noise, border, deletion, patches, grid and clutter, has been used to evaluate the robustness of Recursive Cortical Network [24]. Street view house numbers (**SVHN**), **ICDAR 2013**, and **IIIT5k** [50] are natural text datasets that contain complex scenes with noise and disturbed objects such as fences, bricks and shadows.

We also extend the evaluation on natural images. The **LS-BSDS300** dataset [41] contains various complex scenes and is one of the most difficult datasets in the skeleton detection task. Since unsupervised skeleton detection methods are typically evaluated on LS-BSDS300, our approach is an unsupervised method, LS-BSDS300 is selected as the natural image skeleton dataset for comparison.

Compared methods We compare our approach with six mainly unsupervised skeletonization methods: 1) AMAT [19] considers the detection of the skeleton of natural images as a weighted geometric set covering; 2) Distance transform uses the skiimage.morphology.medial_axis function; 3) Thinning uses the skimage.morphology.thin function; 4) Voronoi uses the scipy.spatial.qhull.Voronoi function; 5) Lindeberg [36] extracts the skeleton by calculating a local maximum or minimum in the main eigendirection of the Hessian matrix, and uses a selection mechanism to determine the scale; 6) Mignotte [40] quantifies the skeleton between each pair of line segments by a Hough-style voting approach and utilizes an averaging procedure to remove noise.

Among these competitors, the AMAT model ranked second in the Skeleton Symmetry Competition in the ICCV 2017 workshop. The first-ranked method namely RSRN [44], is based on a deep learning model, which requires training on a large set of images and also requires the ground-truth skeleton of the images. Considering there is no ground-truth skeletons in the text datasets, this approach is not applicable to this problem. Here, we focus our comparisons on unsupervised skeleton extraction methods.

On natural images, we compare with the mainstream unsupervised skeleton extraction methods [36], [40], [42], [51] and Tsogkas [41], which is a supervised learning-based method. AMAT extracts the skeletons of the foreground and background; hence, it is not suitable for the LS-BSDS300 dataset.

To evaluate the application of our approach, we compare our SF-CNN with CNN on Noisy MNIST, and SF-CRNN with CRNN on CAPTCHA.

Parameters We set the size of the Skeleton Filter as follows: Skeleton Filter size is 31 by 31, Gabor-like filter size is 21 by 21, and Gaussian filter scale is 4. The Gabor-like filter and the Gaussian filter form the internal structure of the Skeleton Filter. Generally, the threshold value of our filter can be set as 0.1. The threshold can be slightly adjusted according to the dataset to yield a superior result. In the skeleton detection experiments, we use the resize parameter to adjust the resolution of the input images. Noisy MNIST uses resize = 4, BotDetect uses resize = 2, Paypal and reCAPTCHA use resize = 2.5. Testing images from IIIT5k are scaled to height 80. The widths are scaled proportionally with the heights.

In the quantitative experiments, the CNN used is the vgg16 model [52]. We set the hyper-parameters as follows: learning rate (0.0002), batch size (100), Adam optimizer, L2 regularization (0.01) on the last fc6 layer and maximal number of iterations (30). The resolution of each train/test image is 224$\times$224 pixels.

The hyper-parameters of the CRNN model [48] are as follows: learning rate (0.01), batch size (64), RMSprop optimizer and maximal number of iterations (200). The resolution of each train/test image is 100$\times$32 pixels.

Implementation details We adopt a simple multi-resolution method to address multi-scale problems in processing arbitrary objects in natural images. The input image uses the CIELab color space, brightness channel L* and color channels a* and b*. For each input image, there are a total of N_{map} output maps.

$$N_{map} = N_{channel} \times N_{resolution} \times N_{orientation} \quad (7)$$

where $N_{channel}$ is the number of input channels, $N_{resolution}$ is the number of resolution levels, and $N_{orientation}$ is the number of orientation channels $\{0, \pi/8, ..., 7\pi/8\}$.

In our experiments, we use features at 4 resolution levels $\{1, 0.5, 0.25, 0.125\}$, 3 input channels (Lab space), and 8 orientation channels for natural images, and we use features at 1 resolution levels, 1 input channels (gray space), and 8 orientation channels for noisy text images. All output maps are combined via element summation operations. For low-resolution maps, we use the bicubic method to upsample to the original image size.

B. Qualitative evaluation on noisy text images

We compare the skeleton detection results on CAPTCHA and natural scene text images. The Skeleton Filter, AMAT,

Fig. 6. Skeleton detection results on CAPTCHA. From left to right: BotDetect (2nd, 3rd columns), reCAPTCHA (4th column) and Paypal (5th column) datasets. Skeleton detection methods include Distance transform, Thinning, Voronoi, Lindeberg's method, Mignotte's method, AMAT and our Skeleton Filter.

Fig. 7. Skeleton detection results on the natural scene text. From left to right: SVHN (2nd and 3rd columns) and ICDAR 2013 (4th and 5th columns) datasets in the natural world. Skeleton detection methods include Voronoi, Lindeberg's method, Mignotte's method, AMAT and our Skeleton Filter. Here, the distance transform and thinning methods are less effective.

Fig. 8. Skeleton detection results on IIIT5k. Skeleton detection methods include Voronoi, Lindeberg's method, Mignotte's method, AMAT and our Skeleton Filter. Here, the distance transform and thinning methods are less effective.

Mignotte's method, Lindeberg's method, distance transform, thinning and Voronoi are tested. Figure 6 presents the skeleton detection results on the BotDetect, reCAPTCHA and Paypal datasets, which demonstrate that our skeleton extraction results outperforms the state-of-the-art approaches.

To reduce the influence of noise on the compared methods, we apply simple post-processing steps to the extracted skeletons. As AMAT extracts the skeleton from both the objects and the background, we use a threshold to filter out the background skeleton and noisy edges. The distance transform method must transform the gray or color images to binary images before extracting the skeletons. A threshold can be set to prune some sub-branches of skeletons. The thinning method must transform the gray or color images to binary images before extracting the skeletons. Voronoi needs to obtain the edge points to compute skeletons; hence, we use the Canny operator to obtain edge points with the default Gaussian smoothing setting. Both Linderberg's and Mignotte's methods have the denosing procedure; thus, we use the default settings of these two methods.

Figure 7 presents the skeleton detection results on the natural scene text. SVHN and ICDAR are typically used for digital and text recognition of natural scenes. To further illustrate the robustness of our approach in noisy conditions, we also conduct experiments on the skeleton extraction of license plate characters in the rain, as shown in the last column in Figure 7. Figure 8 presents the skeleton detection results on IIIT5k. The Skeleton Filter exhibits excellent robustness and outperforms the classical skeletonization methods on the challenging text datasets. The Skeleton Filter is thus a simple but efficient and stable solution for detecting skeletons from the nosiy text images.

C. Quantitative evaluation on the application of skeleton text

To quantitatively evaluate our approach, considering there is no ground-truth skeleton in the noisy text datasets, we compare our SF-CNN pipeline, which classifies outputs of the Skeleton Filter with a standard CNN classifier on Noisy MNIST [24] and CAPTCHA. The Noisy MNIST dataset has three levels of intensity. In the experiments, we selected level 2, which is the noisiest level for testing. Figure 9 presents the skeleton detection results on the Noisy MNIST and MNIST datasets. Our Skeleton Filer outperforms the other approaches, especially for noisy images.

The used CNN is the VGG16 model [52], which was pre-trained on ImageNet, and then fine-tuned at the fc6 fully connection layer with the MNIST dataset. SF-CNN combines Skeleton Filter with the above CNN, where the data are pre-processed by the Skeleton Filter and then classified by the above CNN. SF-CNN and CNN were trained on the 1K and 60K MNIST respectively, and were tested on the 10K Noisy MNIST dataset.

Figure 10 plots the classification accuracy of our SF-CNN with CNN on Noisy MNIST. Compared to CNN, we achieve improvements of 11.6% and 9.5% on average for different types of noise. In Figure 10, for the 1K training sample, our model achieves improvements of 25.9% and 19.0% in background noise and grid scenarios; and for the 60K training sample, our model achieves improvements of 23.1% and 15.1% in background noise and clutter scenarios.

To further verify the effectiveness of the skeletons extracted by our Skeleton Filter, we also conduct a comparative experiment on the recognition of CAPTCHA. The melting-heat set in

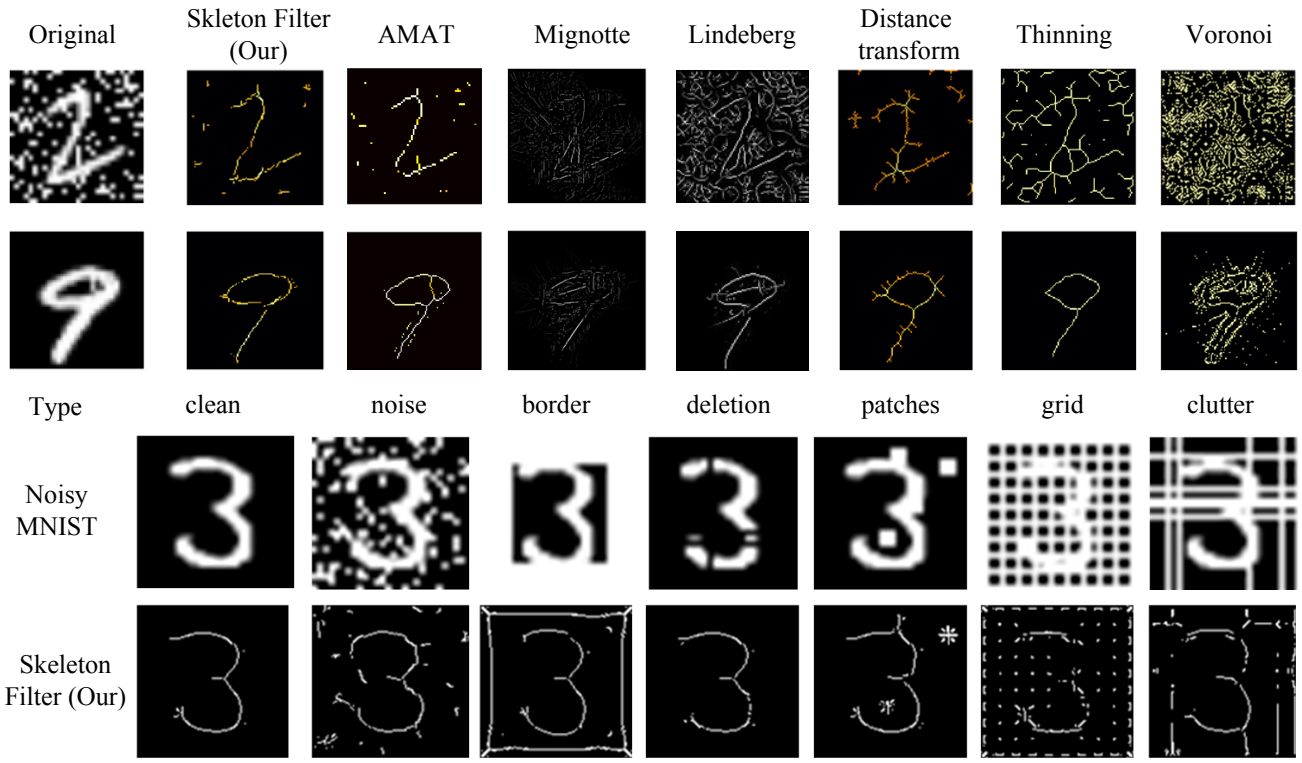
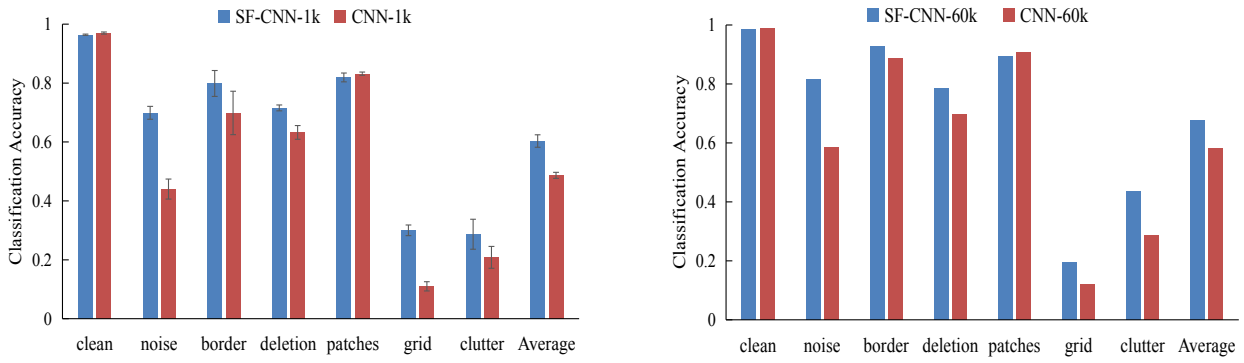


Fig. 9. Skeleton detection results on Noisy MNIST and MNIST. Up: Skeleton detection methods are our Skeleton Filter, AMAT, Mignotte's method, Lindeberg's method, distance transform, thinning and Voronoi; Bottom: Our Skeleton Filter detection results for six variants of noise in the Noisy MNIST dataset.



(a) Training size: 1k

(b) Training size: 60k

Fig. 10. Classification accuracy for SF-CNN and CNN on Noisy MNIST. Legends show the total number of training examples. Testing sizes are 10k.

the Botdetect dataset (seen the 3rd column in Figure 6), a type of CAPTCHA, contains data in which the text and background are of the same color, thereby resulting in melting effects.

We use CRNN [48] as the recognition model on CAPTCHA. CRNN was pre-trained on the Synthetic Word Dataset [53], which consists of 8 million images covering 90K English words, and the network was fine-tuned using the gen-CAPTCHA dataset including 10K images generated by a CAPTCHA module, and the sf-genCAPTCHA dataset which consists of the skeleton data extracted by our method from genCAPTCHA.

In Table I, the baseline is the CRNN-pre model. The recognition accuracy of the baseline model is only 26.0%, although this model has been pretrained on 8 million text images. The CRNN-tune model uses the genCAPTCHA dataset to fine-tune CRNN-pre and test on the melting-heat set of Botdetect dataset. The recognition accuracy of CRNN-tune is 31.0%. SF-CRNN-tune combines Skeleton Filter with the above CRNN-tune, where the data are first processed by Skeleton Filter and then recognized by the above CRNN-tune. The recognition accuracy of SF-CRNN-tune is 53.0%.

According to the results of the two quantitative experiments,

TABLE I
RECOGNITION ACCURACY(%) ON CAPTCHA. CRNN-PRE: CRNN [48] PRETRAINED MODEL ON THE SYNTHETIC WORD DATASET [53] CONTAINING 8 MILLION TRAINING IMAGES; CRNN-TUNE: CRNN-PRE WITH FINE-TUNING ON GENCAPTCHA; SF-CRNN-TUNE (OURS): COMBINES OUR SKELETON FILTER WITH THE ABOVE CRNN-TUNE, WHERE THE FINE-TUNING AND TESTING DATA ARE PREPROCESSED BY THE SKELETON FILTER AND THEN RECOGNIZED BY CRNN-TUNE. THESE MODELS ARE EVALUATED ON THE MELTING-HEAT SET OF OF BOTDETECT DATASET (SEEN THE 3RD COLUMN IN FIGURE 6), A TYPE OF CAPTCHA.

Experiment model	Training sample	Word acc.	Character acc.
CRNN-pre		26.0	76.2
CRNN-tune		31.0	79.0
SF-CRNN-tune		**53.0**	**87.8**

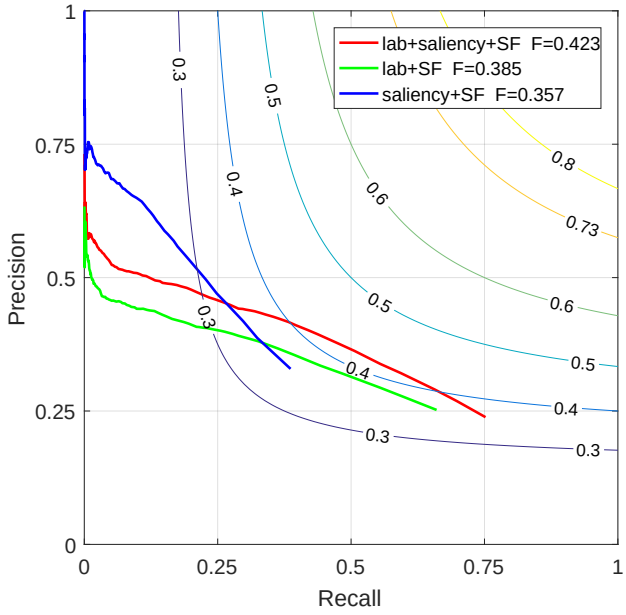


Fig. 11. PR curves for variants of our approach on LS-BSDS300. The skeleton maps are obtained by considering: (1) the Lab image, (2) the saliency feature, (3) merging the skeleton results of cases 1 and 2 via the maximum operation. The saliency feature can reflect the major or large-scale parts of objects, but fail to represent the small-scale parts of objects. The precision of case 2 is high but the recall is relative low. However, the original Lab image preserves all details. It is complementary for cases 1 and 2. Therefore, the skeleton detection performance can be improved by fusing the Lab image and saliency feature.

the skeleton detection results can significantly strengthen the generalization of recognition models, which demonstrates the effectiveness and accuracy of our Skeleton Filter.

D. Qualitative and quantitative evaluation on natural images

We compare our approach with the state-of-the-art skeleton detection methods on LS-BSDS300. Table II lists the F-measure scores for all methods. PR-curves for our methods are plotted in Figure 11. Figure 12 qualitatively compare the proposed approach with other skeleton detectors.

1) For fair comparison, we use the same feature space to detect skeletons. Tsogkas [41] (F-measure = 0.375) uses

TABLE II
F-MEASURES OBTAINED BY THE STATE-OF-THE-ART METHODS AND OUR APPROACH ON LS-BSDS300. LAB, CIELAB COLOR SPACE WHICH INCLUDES BRIGHTNESS AND COLOR FEATURES. CARTESIAN, CARTESIAN SPACE FEATURE. LOG-POLAR, LOG-POLAR SPACE FEATURE. SPECTRAL, SPECTRAL FEATURE. SALIENCY, SALIENCY FEATURE. * INDICATES THAT METHOD IS BASED ON SUPERVISED LEARNING.

Method	Used features **(without texture)**	F-measure
Levinshtein [51]	Superpixel feature + Graph-based clustering	0.356
Lindeberg [36]	Hessian eigenvalues + Automatic scale selecting	0.360
Mignotte [40]	Cartesian + Log-Polar + Hough-style voting	0.362
Tsogkas [41]	Lab + Multiple instance learning	0.375*
SF (our)	Lab + Skeleton Filter	**0.385**
	Used features **(with texture, spectral, saliency)**	
Mignotte [40]	Cartesian + Log-Polar + Texture + Hough-style voting	0.422
Widynski [42]	Lab + Texture + Sequential Monte-Carlo tracking	0.422
Tsogkas [41]	Lab + Texture + Spectral + Multiple instance learning	0.434*
SF (our)	Lab + Saliency + Skeleton Filter	**0.423**

the CIELAB color space to extract brightness and color histogram features, then uses the χ^2-distance to compare the two histograms to predict the probability of symmetric pixels, and finally uses Multiple Instance Learning (MIL) to train their detector to obtain a superior feature vector combinations. Mignotte [40] (F-measure = 0.362) obtains the symmetry feature in the Cartesian and log-polar coordinate space without texture boundary segmentation, and uses a Hough-style voting approach to achieve a better combination. Our Skeleton Filter (F-measure = 0.385) conducts convolution and non-max suppression operations in the CIELAB color space without any prior knowledge. Our approach outperforms the previous methods based on the same feature space.

2) The previous skeleton detection methods are further extended to the other feature spaces to optimize the detection results. Tsogkas [41] (F-measure = 0.434) introduces the texture, spectral clustering features and adopts supervised learning. Mignotte [40] (F-measure = 0.422) adds the texture boundary segments feature in the Cartesian and log-polar coordinate spaces. Widynski [42] (F-measure = 0.422) combines the texture, brightness and color feature, and uses sequential Monte-Carlo to track the symmetric pixels. Our Skeleton Filter (F-measure = 0.423) introduces the saliency features obtained via a saliency detection method [54] without using labels. Our approach can also further introduce these texture and spectral features to optimize the detection performance. However, this paper focuses on solving the skeleton detection task in the noisy text images. For text images, the brightness and color features provide sufficient information for detecting the skeleton.

E. Discussion

Robustly computing the skeleton of objects in noisy images is difficult due to the large variations in the shape boundaries and the large amount of noises in the images. Most previous approaches apply geometric operations to extract the skeleton,

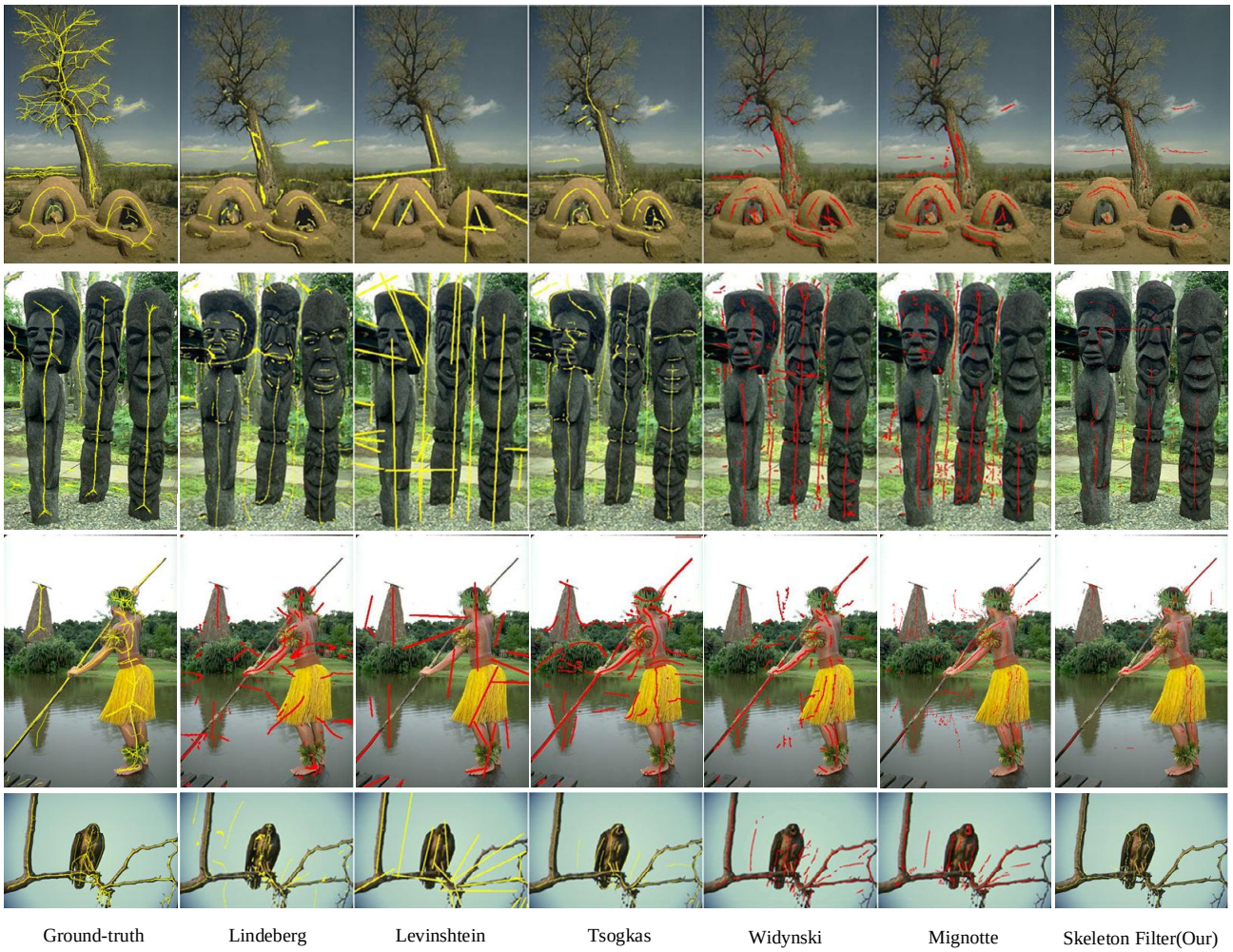


Fig. 12. Qualitative comparison and skeleton detection results against the ground-truth, five other state-of-the-art skeleton detectors (from the second to sixth columns: [36], [51], [41], [42], [40]) and our approach.

which include the distance transform, thinning, Voronoi diagram extraction, graphs, AMAT and etc.. These approaches are highly sensitive to noise and it is difficult to assume that humans conduct such operations in the visual cortex. However, humans can perceive skeletons of objects even under highly noisy conditions. Our Skeleton Filters possess the zero-sum self-symmetric architecture with Gaussian filters. Since Gaussian filter can moderately smooth the image, it can reduce the effects of obvious noise, removing spiky edges that existing skeleton detection approaches suffer from. This zero-sum symmetric architecture can filter out some types of noise distribution. The experimental results demonstrate the robustness of our model.

Skeleton representation perceives a small number of signals from natural objects [7]. This representation has a clear advantage for the invariant shape structure and has been widely used in visual tasks [8], [9]. For example, some research works apply the object skeletons to determine the positional relationships between the corresponding objects.

V. Conclusion

In this paper, we address the problem of perceiving the skeleton of object in dramatically noisy situations. The proposed Skeleton Filter demonstrates excellent robustness and outperforms the classical skeleton extraction methods on the challenging noisy text datasets. The success of Skeleton Filter can inspire researchers in neuroscience to find a similar filter structure within the brain that conducts skeleton extraction.

References

[1] C.-C. Hung, E. T. Carlson, and C. E. Connor, "Medial axis shape coding in macaque inferotemporal cortex," *Neuron*, vol. 74, no. 6, pp. 1099–1113, 2012.

[2] K. Siddiqi, B. B. Kimia, A. Tannenbaum, and S. W. Zucker, "On the psychophysics of the shape triangle," *Vision Research*, vol. 41, no. 9, pp. 1153–1178, 2001.

[3] H. Blum, "A transformation for extracting new descriptors of shape," *Models for the Preception of Speech & Visual Form*, vol. 19, pp. 362–380, 1967.

[4] C. A. Burbeck and S. M. Pizer, "Object representation by cores: Identifying and representing primitive spatial regions," *Vision research*, vol. 35, no. 13, pp. 1917–1930, 1995.

[5] M. Leyton, *A generative theory of shape*. Springer, 2001.

[6] B. B. Kimia, "On the role of medial geometry in human vision," *Journal of Physiology - Paris*, vol. 97, no. 2, pp. 155–190, 2003.

[7] S. M. Pizer, P. T. Fletcher, S. Joshi, A. Thall, J. Z. Chen, Y. Fridman, D. S. Fritsch, A. G. Gash, J. M. Glotzer, and M. R. Jiroutek, "Deformable m-reps for 3d medical image segmentation," *Int J Comput Vis*, vol. 55, no. 2-3, pp. 85–106, 2003.

[8] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[9] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4382–4394, 2018.

[10] A. Shokoufandeh, L. Bretzner, D. Macrini, M. F. Demirci, and S. Dickinson, "The representation and matching of categorical shape," *Computer Vision & Image Understanding*, vol. 103, no. 2, pp. 139–154, 2006.

[11] J. Feldman and M. Singh, "Bayesian estimation of the shape skeleton," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 18 014–18 019, 2006.

[12] M. F. Demirci, A. Shokoufandeh, and S. J. Dickinson, "Skeletal shape abstraction from examples," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 31, no. 5, p. 944, 2009.

[13] P. K. Saha, G. Borgefors, and G. S. D. Baja, "A survey on skeletonization algorithms and their applications," *Pattern Recognition Letters*, vol. 76, pp. 3–12, 2016.

[14] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[15] I. Bitter, A. E. Kaufman, and M. Sato, "Penalized-distance volumetric skeleton algorithm," *IEEE Transactions on Visualization and computer Graphics*, vol. 7, no. 3, pp. 195–206, 2001.

[16] G. Németh, P. Kardos, and K. Palágyi, "Thinning combined with iteration-by-iteration smoothing for 3d binary images," *Graphical Models*, vol. 73, no. 6, pp. 335–345, 2011.

[17] R. Ogniewicz and M. Ilg, "Voronoi skeletons: Theory and applications," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. IEEE, 1992, pp. 63–69.

[18] D. Macrini, K. Siddiqi, and S. Dickinson, "From skeletons to bone graphs: Medial abstraction for object recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[19] S. Tsogkas and S. Dickinson, "Amat: Medial axis transform for natural images," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2727–2736.

[20] P. Li, B. Wang, F. Sun, X. Guo, C. Zhang, and W. Wang, "Q-mat: Computing medial axis transform by quadratic error minimization," *ACM Trans. Graph.*, vol. 35, no. 1, pp. 8:1–8:16, Dec. 2015. [Online]. Available: http://doi.acm.org/10.1145/2753755

[21] F. Chavane, D. Sharon, D. Jancke, O. Marre, Y. Frégnac, and A. Grinvald, "Lateral spread of orientation selectivity in v1 is controlled by intracortical cooperativity," *Frontiers in Systems Neuroscience*, vol. 5, p. 4, 2011.

[22] B. Scholl, A. Y. Tan, J. Corey, and N. J. Priebe, "Emergence of orientation selectivity in the mammalian visual pathway," *Journal of Neuroscience*, vol. 33, no. 26, pp. 10 616–10 624, 2013.

[23] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.

[24] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, and H. Wang, "A generative vision model that trains with high data efficiency and breaks text-based captchas," *Science*, vol. 358, no. 6368, p. eaag2612, 2017.

[25] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, and et al., "Supplementary materials for a generative vision model that trains with high data efficiency and breaks text-based captchas," *Science*, vol. 358, no. 6368, 2017.

[26] T. Poggio and E. Bizzi, "Generalization in vision and motor control," *Nature*, vol. 431, no. 7010, p. 768, 2004.

[27] K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker, "Shock graphs and shape matching," *International Journal of Computer Vision*, vol. 35, no. 1, pp. 13–32, 1999.

[28] J. Feldman, M. Singh, E. Briscoe, V. Froyen, S. Kim, and J. Wilder, *An Integrated Bayesian Approach to Shape Representation and Perceptual Organization*. Springer London, 2013.

[29] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 209–216. [Online]. Available: https://doi.org/10.1145/258734.258849

[30] J. M. Thiery and T. Boubekeur, "Sphere-meshes:shape approximation using spherical quadric error metrics," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–12, 2013.

[31] A. Sud, M. Foskey, and D. Manocha, "Homotopy-preserving medial axis simplification," in *Proceedings of the 2005 ACM Symposium on Solid and Physical Modeling*, ser. SPM '05. New York, NY, USA: ACM, 2005, pp. 39–50. [Online]. Available: http://doi.acm.org/10.1145/1060244.1060250

[32] F. Chazal and A. Lieutier, "The "λ-medial axis"," *Graphical Models*, vol. 67, no. 4, pp. 304–331, 2005.

[33] B. Miklos, J. Giesen, and M. Pauly, "Discrete scale axis representations for 3d geometry," in *ACM SIGGRAPH 2010 Papers*, ser. SIGGRAPH '10. New York, NY, USA: ACM, 2010, pp. 101:1–101:10. [Online]. Available: http://doi.acm.org/10.1145/1833349.1778838

[34] N. Faraj, J.-M. Thiery, and T. Boubekeur, "Progressive medial axis filtration," in *SIGGRAPH Asia 2013 Technical Briefs*. ACM, 2013, p. 3.

[35] F. Sun, Y.-K. Choi, Y. Yu, and W. Wang, "Medial meshes for volume approximation," *arXiv preprint arXiv:1308.3917*, 2013.

[36] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 117–156, 1998.

[37] J.-H. Jang and K.-S. Hong, "A pseudo-distance map for the segmentation-free skeletonization of gray-scale images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 18–23.

[38] Z. Yu and C. Bajaj, "A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion," in *null*. IEEE, 2004, pp. 415–420.

[39] C. Direkoglu, R. Dahyot, and M. Manzke, "On using anisotropic diffusion for skeleton extraction," *International journal of computer vision*, vol. 100, no. 2, pp. 170–189, 2012.

[40] M. Mignotte, "Symmetry detection based on multiscale pairwise texture boundary segment interactions," *Pattern Recognition Letters*, vol. 74, pp. 53–60, 2016.

[41] S. Tsogkas and I. Kokkinos, "Learning-based symmetry detection in natural images," in *European Conference on Computer Vision*. Springer, 2012, pp. 41–54.

[42] N. Widynski, A. Moevus, and M. Mignotte, "Local symmetry detection in natural images using a particle filtering approach," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5309–5322, 2014.

[43] W. Shen, K. Zhao, Y. Jiang, Y. Wang, Z. Zhang, and X. Bai, "Object skeleton extraction in natural images by fusing scale-associated deep side outputs," in *Computer Vision and Pattern Recognition*, 2016, pp. 222–230.

[44] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye, "Srn: Side-output residual network for object symmetry detection in the wild," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 302–310, 2017.

[45] C. Liu, W. Ke, J. Jiao, and Q. Ye, "Rsrn: Rich side-output residual network for medial axis detection," in *IEEE International Conference on Computer Vision Workshop*, 2017, pp. 1739–1743.

[46] T. Sing Lee, D. Mumford, R. Romero, and V. Lamme, "The role of the primary cortex in higher level vision," *Vision research*, vol. 38, pp. 2429–54, 09 1998.

[47] J. Keat, P. Reinagel, R. C. Reid, and M. Meister, "Predicting every spike: a model for the responses of visual neurons," *Neuron*, vol. 30, no. 3, pp. 803–817, 2001.

[48] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[49] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in *Advances in Cryptology — EUROCRYPT 2003*, E. Biham, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 294–311.

[50] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *BMVC*, 2012.

[51] A. Levinshtein, C. Sminchisescu, and S. Dickinson, "Multiscale symmetric part detection and grouping," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 117–134, 2013.

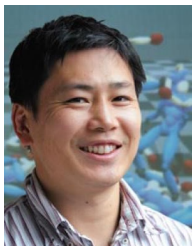
[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.
[53] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.
[54] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *ECCV*, 2018.

Xiuxiu Bai received the B.S. degree from Xi'an Jiaotong University in 2009, and the Ph.D. degree from Xi'an Jiaotong University in 2016. She visited Edinburgh University, from 2017 to 2018. She is currently an Assistant Professor with the School of Software Engineering, Xi'an Jiaotong University. Her research interests include computer vision and visual neuroscience.

Lele Ye received his B.S. in Software Engineering from XinJiang University, Urumqi, China, in 2017. He is currently a master student of the the School of Software Engineering, Xi'an Jiaotong University. His research interests include computer vision and machine learning.

Jihua Zhu received the B.E. degree in automation from Central South University, China, and the Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China, in 2004 and 2011, respectively. He is currently an Associate Professor with the School of Software Engineering, Xi'an Jiaotong University. His research interests include computer vision and machine learning.

Li Zhu received the B.S. degree from Northwestern Polytechnical University in 1989 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University in 1995 and 2000, respectively. He is currently an Associate Professor with the School of Software, Xi'an Jiaotong University. His main research interests include multimedia processing and parallel computing.



Taku Komura is a Reader (Associate Professor) at the Institute of Perception, Action and Behaviour, School of Informatics, Edinburgh University. He is also a Royal Society Industry Fellow. He received his B.Sc, M.Sc. and D.Sc in Information Science from the University of Tokyo. His research interests include character animation, computer graphics and interactive techniques.